Short Segmental Duplication: Model for Growth of Microbial Genomes

Li-Ching Hsieh*, Liaofu Luo† and H.C. Lee*§
*Department of Physics and §Department of Life Science,
National Central University, Chungli, Taiwan 320
†Department of Physics,
University of Inner Mongolia Hohhot 010021, China

(Dated: April 15, 2004)

We show that textual analysis of microbial complete genomes reveals telling footprints of their early evolution. If a DNA sequence considered as a text in its four bases is sufficiently random, the distribution of frequencies of words of a fixed length from the text should be Poissonian. We point out that in reality, for words less than nine letters complete microbial genomes universally have distributions that are uniformly many times wider than those of corresponding Poisson distributions. We interpret this phenomenon as follows: the genome is a large system that possesses the statistical characteristics of a much smaller random system, and certain textual statistical properties of genomes observable now are remnants of those of their ancestral genomes, which were much shorter than genomes today. This interpretation motivates a simple biologically plausible model for the growth of genomes: the genome first grew randomly to an initial length of not more than one thousand bases (1 kb), thereafter mainly grew by random segmental duplications. Setting the lengths of duplicated segments to average around 25b, we have generated model sequences *in silico* whose statistical properties emulate those of present day genomes. The small size of the initial random sequence and the shortness of the lengths the duplicated segments both dictate an RNA world at the time growth by duplication began. Growth by duplication allowed the genome repetitive use of hard-to-come-by codes increasing thereby the rates of evolution and species diversion enormously.

Keywords: Genome analysis, statistical properties, evolution, RNA world, genome growth model

I. FREQUENCY OF OCCURRENCE OF OLIGONUCLEOTIDES IN MICROBIAL GENOMES

It is a general rule of statistics that the larger the system the more sharply defined its average properties. When apples are randomly dropped into barrels, the distribution of apples in the barrels is governed by the Poisson distribution. If 1,024 apples were dropped into sixty-four barrels, there is a 5% chance that one of the barrels would have less than 8 or more than 24 apples. If 1 million apples were dropped into the barrels the chances that the number of apples received by any barrel fall outside the range of 14,600 to 16,600 would be exceedingly small, and there is a less than one in 10^{980} (10^{830}, respectively) chance that one barrel would get as few (many) as 8,000 (24,000) apples.

Microbial genomes are seemingly random systems when viewed as texts of the four bases represented by A, C, G and T. To count the number of times each of the sixty-four trinucleotides, or 3-mers, occur in a genome-as-text is similar to counting the number of apples after they have been dropped into barrels. The genome of the bacterium *Treponema pallidum*, the causative agent of syphilis, is about 1M base pairs long and has almost even base composition [1]. In an astonishing departure from what is expected of a system of its size, the genome has six 3-mers (CGC, GCG, AAA, TTT, GCA, TGC) occurring more than 24,000 times per 1 Mb and two (CTA, TAG) less than 8,000 times. Scrambling the genome sequence thoroughly restores it to a random sequence obeying Poisson distribution and the large-system rule.

T. pallidum is not exceptional in disobeying the large-system rule. For the twenty-five complete microbial "Class A" genomes whose combine probability p for AT or CG content is 0.46 to 0.55, the observed standard deviation (std) of the distribution of the frequency of occurrence of 3-mers per 1 Mb (hereafter called 3-distribution) is 4,080±630 around the mean of 15,625. This is about 32 times the std of a Poisson distribution of the same mean that a random sequence would yield.

TABLE I: For given k's, standard deviation of k-distributions: for the genome *T. pallidum*; averaged over 25 Class A genomes; for a random sequence; for a Class A model sequence (see text). The last column gives the length (L_{stat}) of a random sequence with the genomic ratio of mean count to std. The figures following ± in the third column give the stds associated with the average stds. All stds are normalized to correspond to a base composition of 50% AT (see Methods).

k	*T. pal*	Class A	Ran.	Model	L_{stat} (nt)
2	8227	10610±2107	250	8207	.65±.35
3	3977	4379±707	125	3415	1.0±0.3
4	1384	1490±232	62.5	1202	1.9±0.5
5	434	468±72.5	31.2	402	4.7±1.3
6	129	141±22.3	15.6	134	13±4
7	37.5	41.6±7.0	7.8	45.3	37±12
8	11.0	12.3±2.3	3.9	15.9	110±40
9	3.4	3.76±0.85	1.9	5.9	300±130
10	1.3	1.29±0.34	1.0	2.3	640±300

Nor is the statistics of 3-mers special in genomes. In Table I, column 3 gives the std of the k-distribution, k = 2 to 10, averaged over the twenty-five Class A genomic sequences and column 4 gives the std for a Poisson distribution with mean value $10^6/4^k$. The genomic stds approach those of a random sequence when k increases beyond 10. For k less than 10, the Poisson std increases as 2^{-k} with decreasing k whereas the genomic std increases at a much higher rate, such that already at k=8 the genomic std is many times greater than the Poisson std. Because the variance in the ge-

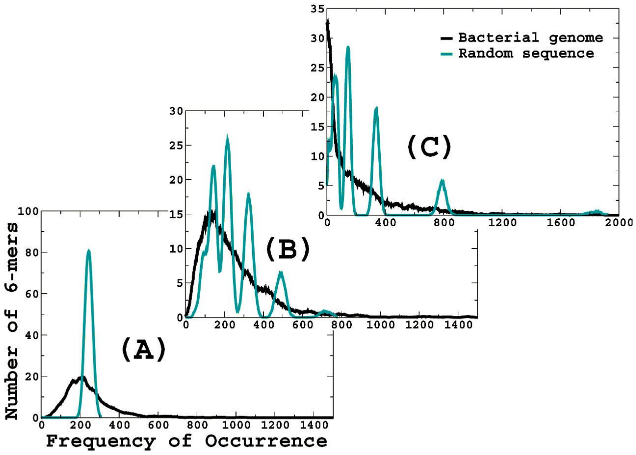

FIG. 1: Comparison of 6-distributions of genomes (black) and random sequences (green/gray), with abscissa giving the frequency of occurences of 6-mers and ordinates showing the number of 6-mers having a given frequency: (A) *T. pallidum* genome and a random sequence with $p = 0.5$; (B) *C. muridarum* genome and a random sequence with $p = 0.6$; (C) *M. jannaschii* genome and a random sequence with $p = 0.7$

nomic std is typically much smaller than the difference between the genomic and Poisson stds, the genomic k-distribution differs from the Poisson distribution in a universal fashion. Hence we shall speak of a universal (Class A) genome.

The base composition of a genome has a conspicuous effect on its k-distributions. The black curve in Figure 1 shows 6-distributions of three representative genomes: (A) *T. pallidum* with $p \approx 0.5$; (B) *Chlamydia muridarum* [2] with $p \approx 0.6$; (C) *Methanococcus jannaschii* [3] with $p \approx 0.7$. For comparison the green/gray curves in the figure show 6-distributions of random sequences with $p = 0.5$, 0.6 and 0.7, respectively. The k-distribution of a random sequence is composed of k+1 Poisson distributions with mean frequencies $10^6\ 2^{-k}p^m(1\text{-}p)^{k-m}$, m=0 to k, which coalesce into a single Poisson distribution when p is close to 0.5. In contrast, narrow sharp spikes are completely absent in the 6-distributions for the microbial genomes. In what follows we leave aside whatever complication large differences in base composition may cause and first focus our attention on Class A genomes.

II. MICROBIAL GENOMES ARE LARGE SYSTEMS WITH SMALL-SYSTEM STATISTICS

The (Class A) universal genome has the statistical property of a random sequence much shorter in length than itself. To see this, we define the "statistical length" L_{stat} of the universal genome as the length of a random sequence that has a k-distribution with a mean to std ratio equal to that of the corresponding genomic ratio r. Then L_{stat}=4^kr^2, and its values for the various k's are given in the last column of Table I. L_{stat} has a strong k dependence: it is very short for the smaller k's - of the order of 1 kb for k≤3 - and grows rapidly with k. When k=10 it is about half the length (normalized to 1 Mb) of the real genome.

A signature of the universal genome, by comparison to a random sequence, lies in its very large numbers of both overrepresented and underrepresented oligonucleotides. As a typical representative of the universal genome, the genome of *E. coli* [4] has 500 and 510 6-mers whose frequencies of occurrence are greater than 400 and less than 100 per 1 Mb, respectively, while a random sequence has none in either category. There are many known examples of individual oligonucleotide that exhibit extreme relative abundance. For dinucleotides this was noted to be common and has genome-wide consistency [5]; tetra- and hexapalindromes are almost always underrepresented in bacteriophages and are systematically underrepresented in bacteria where 4-cutting and/or 6-cutting restriction enzymes are common [6]; an 8-mer that appears as Chi sites, hotspots of homologous recombination, is highly overrepresented in *E. coli* [7]; in the human pathogens *Haemophilus influenzae* [8, 9] and *Neisseria* [10] there are 9- and 10-mers functioning as uptake signal sequences that are vastly overrepresented. The causes for these extreme cases are generally not known and, with the exception of the dinucleotides, such individual cases do not decisively determine the statistical properties of a genome.

What caused a genome to have k-distributions so much wider than those of a random sequence? Natural selection suggests itself as a prime explanatory candidate. For instance, the 64 frequencies of codons, 3-mers used by the genome to code proteins in genes, exhibit very wide distributions. But natural selection by itself does not directly cause any change in a genome. Such changes are caused by random mutations and other stochastic mechanisms. Natural selection may account for what changes come to pass; if, however, such changes always tend to promote or retain a randomness that exhibits Poisson distribution, then the ability of natural selection to push the genome very far in a non-Poissonian direction would seem to have its limits.

III. MODEL FOR EARLY GENOME GROWTH

Here we propose a biologically plausible model for the growth and evolution of a universal genome that can generate the observed statistical characteristics of genomic sequences. The model is very simple and consists of two phases. In the first phase the genome initially grows to a random sequence whose size is much smaller than the final size of the genome. In the second phase the genome grows by random segmental duplications possibly modulated by random single mutations. In this work a snapshot is taken of the model sequence shortly after it reaches a length of 1 Mb. The key aspect of the model is growth by segmental duplication, the most straightforward and biologically viable way for the universal genome to become what it appears to be - a large system that exhibits small-system statistical characteristics.

Growth by whole-genome duplication [11–13] coupled with mutation is ruled out because such a mode of growth yields genomes whose k-distributions have the incorrect k-dependence - their L_{stat} vary with k

too weakly. Indeed we found it comparatively easy to generate a sequence that could faithfully reproduce the genomic k-distribution for any given k, say $k = k'$, but not simultaneously those of other k's. Typically such a sequence had an L_{stat} that has a k-dependence far too weak than required to fit genomic data and, consequently, k-distributions that are too narrow when $k < k'$ and too broad when $k > k'$. Ways to generate several such examples are given in the Methods. Generating a sequence that would emulate a real genome was a much more exacting task.

IV. METHODS

Complete microbial genome sequences are obtained from GenBank [14]. The names and GenBank codes of the 25 Class A genomes ($p \approx 0.5$) and the class of 28 $p \approx 0.7$ genomes used in our analysis are given in the Supporting Information. The codes for *T. pallidum*, *M. jannaschii* and *C. muridarum* are NC_000919, NC_000909 and NC_002620, respectively. Counting of k-mers is done by reading through a k-base wide sliding window. Counts are normalized to per 1 Mb and variation in genomic base composition is compensated for by dividing the actual genomic counts by the factor $L(p/\bar{p})^m((1-p)/(1-\bar{p}))^{k-m}$, where p is the AT (or CG, whichever is more numerous) content of the genome, $\bar{p} = 0.5$ ($\bar{p} = 0.7$) for Class A (the class of $p = 0.7$) genomes and m is the total number of AT bases (or CG) in each k-mer.

Generation of model sequence. A random sequence of length L_0 with a given base composition is first generated. Thereafter the sequence is altered by single mutations (replacements only) and duplications, with a fixed average mutation to duplication event ratio. In duplication events, a segment of length l, chosen according to the Erlang probability density function $f(l) = 1/(\sigma n!)(l/\sigma)^n e^{-l/\sigma}$, is copied from one site and pasted onto another site, both randomly selected. In the above, n is an integer and σ is a length scale in bases. The function gives a mean duplicated segment length $\bar{l} = (n+1)\sigma$ with std $\delta_l = (n+1)^{1/2}\sigma$. The values $n = 0$ to 8 and selected values for σ from 3 to 15,000 were used. In the text, the model sequences used to compare with genomic sequences were generated with $L_0 = 1000$, $n = 4$, $\sigma = 5$ and without mutation events. This model has $\bar{l} = 25$ and $\delta_l = 11.2$. When $f(l)$ is replaced by a Gaussian distribution with the same values for $\bar{l}$ and δ_l, respectively, less satisfactory results are obtained. Fine-tuning to find the best parameters was not attempted. The following are some examples that gave very good k-distributions for specific k-mers but not generally; all were generated with $L_0 = 1000$ and $n = 0$: for 6-mer, $\sigma = 13,000 \pm 2,000$ and on average 0.04σ mutations per duplication (these parameters also work for genomes with biased base compositions) [15]; for 2-mer, $\sigma = 50$, no mutation; for 5-mer, $\sigma = 30$, no mutation; for 9-mer, $\sigma = 15$, no mutation.

Presentation of data. In Figs. 3 and 5 the curves shown are the result of a small amount of forward and backward averaging - to remove excessive fluctuations. In Figs. 2 and 4 data bunching was used to produce the towers shown.

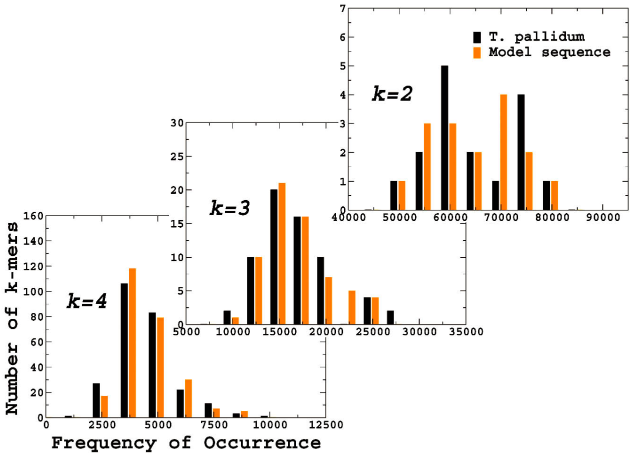

FIG. 2: Histograms of k-distributions for genome of *T. pal.* (black) and a Class A ($p = 0.5$) model sequence (orange/dark gray), k=2 to 4, with abscissa indicating intervals of frequency of occurrence of k-mers and ordinates giving the number of k-mers falling within a given interval of frequency of occurrence. In each case the histogram of the k-distributions for a random sequence would be represented by a single tower located at the mean frequency $10^6 4^{-k}$.

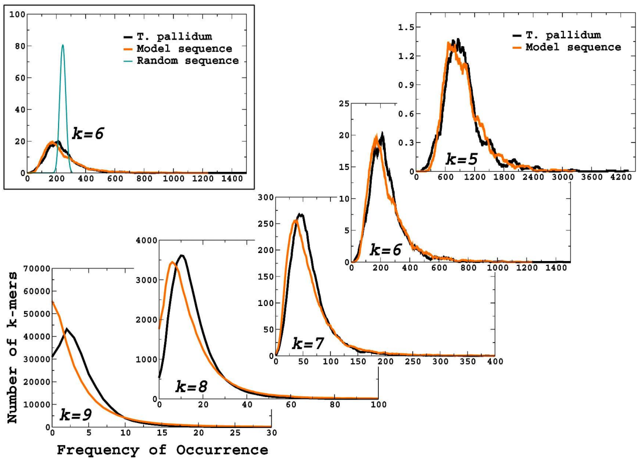

FIG. 3: k-distributions for genome of *T. pal.* (black) and a Class A ($p = 0.5$) model sequence (orange/dark gray), k= 5 to 9. See legend of Fig. 1 for further detail. The top-left panel shows 6-distributions for *T. pallidum* the model sequence and a random sequence (green/gray) with $p = 0.5$.

V. RESULT

After extensive experimentation, it was found that sequences having the statistical characteristics sought after could be generated by choosing: (i) the length (L_0) of the initial random sequence to be approximately 1 kb; (ii) the average length ($\bar{l}$) of the (randomly chosen) duplicated segments to be 25b with a spread (δ_l) of approximately 11b. It is emphasized that every step in the growth procedure is taken stochastically.

The stds of the k-distributions of a model sequence thus generated are given in column five of Table I. They agree quite well with the observed genomic values in columns two and three although their k-dependence is still slightly too weak. Histograms of k-distributions of *T. pallidum* (black) and the model sequence (orange/dark gray), k=2, 3 and 4, are compared in Fig. 2. In all three cases, the histogram for a random sequence would be represented by a single narrow tower located at the mean frequency. For k=2 and to a lesser extent k=3, the histograms for both genomic and model sequences display large fluctuations. The model sequence is not expected to exactly reproduce the counts of the genomic sequence. Indeed, generated stochastically, another model sequence (generated using the same parameters) will yield histograms that differ in detail from those shown in the k=2 and 3 panels of Fig. 2 but show patterns of fluctuation still similar to those exhibited by the genomic sequence and have stds very close to those given in column 5 of Table I. Fig. 3 shows comparisons for k=5 to 9. The panel in the top-left corner gives the 6-distributions from *T. pallidum*, a random sequence (green/gray) and the model sequence. In every case the model sequence succeeds in broadening out the narrow peaks that come with a random sequence and has k-distributions very similar to those obtained from *T. pallidum*.

The same growth model also can account for the k-distributions of genomes with biased base composition. We demonstrate this by comparing the k-distributions for *M. jannaschii*, which has $p \approx 0.7$, with those of a model sequence in Figs. 4 and 5. The model sequence was generated using exactly the same procedure and parameters described above for generating the Class A model sequence, with the only exception being that the initial 1 kb random sequence has $p = 0.7$ rather than 0.5. The distributions for k=2, 3 and 4 are shown in Fig. 4 as histograms and those for k=5 to 9 are shown in Fig. 5. The top-left panel of Fig. 5 reminds us that a k-distribution from a random sequence with p significantly different from 0.5 has k+1 narrow peaks, which are wholly absent from the k-distributions of both the *M. jannaschii* and the model sequence. Our general remarks concerning the comparison of k-distributions of *T. pallidum* and the Class A model sequence also apply here. Suffice to say that in all cases the stochastically generated model sequence succeeds in reproducing key features of the k-distributions of the *M. jannaschii* genome.

The k-dependent statistical lengths L_{stat} for genomes with highly biased based compositions can no longer be extracted from the overall widths, but rather from widths of distributions of subsets of k-mers with fixed AT content (see Supporting Information). From 28 microbial genomes with $0.66 \leq p \leq 0.75$ we obtain for L_{stat} (in k nt): 0.53±0.30, 1.1±0.6, 2.1±1.1, 5.2±2.5, 14±6, 36±17, 93±44, 230±110, 600±240, for k = 2 to 10, respectively. These values are in very good agreement with the L_{stat}'s given in the last column in Table I. This strengthens the notion of universality that microbial genomes are large systems with small-system statistics. To summarize, we have two pieces of evidence suggesting that genomes with widely varying base compositions and content have essentially the same kind of growth histories, whereby (a) they have essentially identical sets of L_{stat}'s and (b) their respective k-distributions are emulated by model sequences generated using identical parameters (except in the base composition of the initial random sequence). The implication is that the only significant difference between the two classes of genomes is in their base composition, and this difference was substantially in place before the genomes began to grow by duplication.

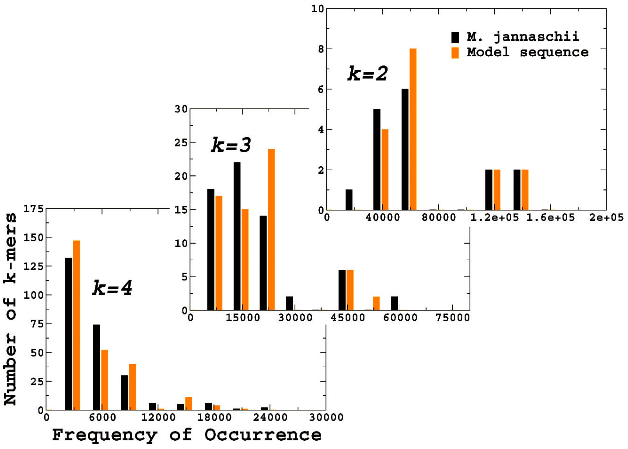

FIG. 4: Histograms of k-distributions for genome of *M. jan.* (black) and a p = 0.7 model sequence (orange/dark gray), k=2 to 4. See legend of Fig. 2 for further detail. The histograms for a random sequence would be given by $k + 1$ narrow towers.

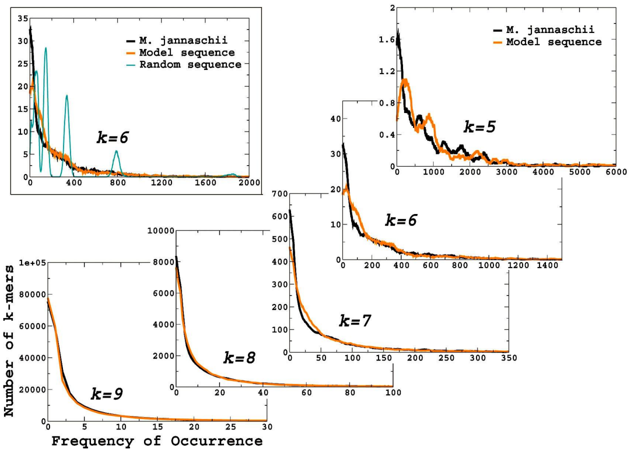

FIG. 5: k-distributions for genome of *M. jan.* (black) and a p = 0.7 model sequence (orange/dark gray), k= 5 to 9. See legend of Fig. 1 for further detail. The top-left panel shows 6-distributions for *M. jannaschii*, the model sequence and a random sequence (green/gray) with p = 0.7.

The model sequences are parameter-sensitive. If L_0 was significantly longer than 1 kb then no good model sequence could be found. This is unsurprising because L_0 cannot be much longer than the shortest L_{stat} in Table I. If, with L_0=1 kb, either $\bar{l}$ or δ_l was changed by more than 10% from their optimal values of 25b and 11b respectively then the agreement between the genomic and model sequences would worsen noticeably.

VI. DISCUSSION

In the spirit of simplicity no (point) mutations were imposed on the model sequences whose properties are shown here. This is because mutations randomize a sequence whereas we are looking for a way to stochastically generate model sequences with less randomness than that of a random sequence. If the model sequence were allowed to have too many mutations then a compensating mechanism would have to be found to derandomize the sequence. In reality the randomizing effect of mutation is partially compensated for by the stabilizing force of natural selection, whose effect is also not explicitly considered in our model. In any case the model sequences as given can tolerate about one mutation per two duplication events: twenty thousand mutation fixations reduces the std of the k-distributions of the Class A model sequence by 4% (for k=2) to 10% (for k=10).

In bacterial genomes, typically about 12% of genes represent recent duplication events - 12% in *T. pallidum* [1], 11.2% in *H. influenzae* [16] and 12.8% in *V. cholerae* [17]. The model sequences presented here do not explain the pattern of all such duplications, many of which would involve segments up to several kb long. Work is underway to extend the model to account for the genomic pattern of repeat sequences of all lengths.

The generic statistical textual properties of eukaryotic genomes have also been examined and findings will be reported elsewhere. So much now is believed to obtain: when the great difference in length between microbial and eukaryotic genomes is accounted for, what is said here of the statistical textual properties of microbial genomes should hold true *mutatis mutandis* for eukaryotic genomes.

To be sure there will be many textual aspects of the generic microbial genome that the growth model proposed here will not be able to account for in detail. Nevertheless we believe the evidence presented here is sufficiently strong to support the following proposition: the ancestors of microbial genomes underwent a fundamental transition in their growth and evolution shortly after they had reached a length of not more than 1 kbp and had acquired a rudimentary duplication machinery, thereafter grew (and diverged) mainly by stochastic duplication of short segments whose lengths averaged to about 25b. Assuming this model to be substantially correct we mention some of its implications for biology and evolution.

Our results suggest that the base composition of a genome was essentially inherited from an ancestor whose own composition had been determined either randomly or by some unknown cause by the time of the transition to growth-by-duplication and that subsequent compositional changes caused by natural selection have been minor. If so, the base composition of a present-day genome should be close to being uniform over the entire genome because of the relatively small size of the ancestor genome - regardless of how it came into being - by comparison to present-day genomes. Indeed, the base composition is known to be essentially uniform over the entire genome [5] (although coding regions often have a relatively richer GC content [3, 16, 18, 19]) and this phenomenon is of unknown biological significance.

A genome of the order of 1 kbp long is far too short to encode enough proteins for DNA duplication. Setting the initial length of our model universal genome to not greater than 1 kbp at the point of transition to growth-by-duplication thus necessarily implies that the universal genome began its life in an RNA world [20, 21], when there were no proteins and when RNAs had the dual roles of genotype and phenotype (see [22] for a review). This view of the origin of life was advocated [23–25] even before RNA was discovered to exhibit self-splicing and enzymatic activities [26, 27]. Some RNA enzymes, or ribozymes, are very small; the hammerhead ribozyme is only 31 to 42 nucleotides (nt) long [28] and the hairpin ribozyme is only 50 nt long [29]. Hence we can infer with reasonable certainty that the 1 kbp initial universal genome was of sufficient size to encode the machinery necessary for sustained evolution and duplication. Our model does not address the origin of this initial genome. The likelihood that it evolved from something arising spontaneously beforehand is enhanced by its short length and supported by the successful isolation of artificial ribozymes from pools of random RNA sequences *in vitro* [30]. The average duplicated segment length of 25b likely represents a good portion of the length of a typical ribozyme encoded in the early universal genome even if it is very short compared to a present-day gene that codes for an enzyme.

Natural selection has been largely ignored in our model not because it is unimportant in evolution, but rather because we believe it could not have played a lead role in generating the non-Poissonian statistical characteristics of genomes discussed here. Suppose the genome grew not by segmental duplication but via "events" in which oligonucleotides were inserted at random into the genome (event specifics are unimportant to our argument) and at every such event the oligonucleotide was accepted or rejected through natural selection according to some preference causing a broadening of the k-distributions of the genome. If the oligonucleotides had been random, then there would have been many rejections between acceptions and the number of events needed to generate a genome with k-distributions as wide as those of real genomes would have been orders of magnitude greater than the number of segmental duplications required to achieve the same effect. Against a scenario in which change was driven by natural selection, the principle of parsimony would dictate that segmental duplication was the overwhelming dominant force generating the wide k-distributions we now observe.

As a corollary of the above reasoning, we can consider that uneven codon usage may not have been the primary cause of the very broad distribution of the 3-mer counts now seen in genomes. It is much more likely that codons were evolutionary "spandrels" [31], that is to say, their rise as codes for proteins came as a consequence of an opportunistic evolutionary adaptation to the already-wide 3-distribution that had resulted from growth by duplication. Similarly, many of the highly under- or overrepresented oligonucleotides that now have biological functions might have originated as spandrels.

Our analysis presented here supports the view that statistical characteristics of present day genomes were already substantially determined by the characteristics of their ancestors by the time of their transition to growth by duplication; conversely, that statistical characteristics of genomes today can be regarded just in this manner as a basis from which to explore the nature and properties of early ancestral genomes; further analysis made along this line of reasoning may bring us a step nearer in understanding the universal ancestor [32].

Growth by duplication is in itself a brilliant strategy because it allowed the genome to utilize hard-to-come-by codes repeatedly, thereby increasing the rates of evolution and species diversion enormously. For this strategy to have worked, the length of the duplicated segments used and the typical length of coding sequences must match. This condition is likely met by our model because most ribozymes in the early universal genome must have been small. Was this strategy continued after the "early life" of the universal genome - after the rise of codons and proteins? If so, then, with proteins/enzymes much larger than the small ribozymes, the duplicated segments in the post-protein era must have been much longer than 25b for the strategy to have been effective. In higher organisms many repeat sequences with lengths ranging from 1 base to many kilobases are believed to have resulted from at least five modes of duplication, and about 50% - perhaps even more - of the human genome is composed of such duplications [18, 19]. Furthermore, as already mentioned, typically about 12% of genes in bacterial genomes represent recent duplication events. So certainly, the continuity of this strategy into the protein era is abundantly in evidence. In eventual answers to questions such as why genes have been duplicated [33] at the high rate of about 1% per gene per million years [34], and why in all life forms so many duplicate genes are found [35–37], this growth strategy, if adopted universally in genomes, may be a simple and crucial, indeed a parsimonious part.

HCL thanks the National Science Council (ROC) for the grant NSC 91-2119-M-008-012, members of the Redfield Lab and the Otto Lab, Department of Zoology, University of British Columbia, Institute for Theoretical Physics, Chinese Academy of Science, Beijing, and the Center for Theoretical Biology, Beijing University, for comments and discussions, and C. S. Lee for editorial comments.

Correspondence and requests for material should be addressed to HCL at hclee@phy.ncu.edu.tw.

[1] C. M. Fraser, *et al.*, Science **281** (1998) 375.
[2] T. D. Read, *et al.*, Nucl. Acids Res. **28** (2000) 1397.
[3] C. J. Bult, *et al.*, Science **273** (1996) 1058.
[4] F. R. Blattner, *et al.*, Science **277** (1997) 1453.
[5] S. Karlin and C. Burge, Trends in Genetics **11** (1995) 283.
[6] S. Karlin, *et al.*, Nucl. Acids Res. **20** (1992) 1363.
[7] T. Colbert, A. F. Taylor and G. R. Smith, Trends in Genetics **14** (1998) 485.
[8] H. O. Smith, *et al.*, Science **269** (1995) 538.
[9] S. Karlin, J. Mrazek and M. Campbell, Nucl. Acid Res. **24** (1996) 4263.
[10] H. O. Smith, *et al.*, Res. Microbiol. **150** (1999) 603.
[11] S. Ohno, *Evolution by gene duplication*, (Springer Verlag, New York, 1970).
[12] L. Skrabanek and K. H. Wolfe, Cur. Op. Gen. and Dev. **8** (1998) 694.
[13] A. L. Hughes, J. da Silva, and R. Friedman, Genome Res., **11** (2001) 771.
[14] GenBank: *www.ncbi.nlm.nih.gov/PMGifs/Genomes /micr.html.*
[15] L. C. Hsieh, Liaofu Luo, Fengmin Ji and H. C. Lee, Phys. Rev. Lett. **90** (2003) 018101.
[16] Arabidopsis Genome Initiative, Nature **408** (2000) 796.
[17] J. F. Heidelberg, *et al.*, Nature **406** (2000) 477.
[18] E. S. Lander, *et al.*, Nature **409** (2001) 860.
[19] J. C. Venter, *et al.*, Science **291** (2001) 1304.
[20] W. Gilbert, Nature **319** (1986) 618.
[21] J. E. Darnell and W. F. Doolittle, Proc. Natl. Acad. Sci. USA **83** (1986) 1271.
[22] G. F. Joyce, Nature **418** (2002) 214.
[23] C. Woese, *The Genetic Code*, (Harper & Row, New York, 1967) 179.
[24] F. H. C. Crick, J. Mol. Bio. **38** (1968) 367.
[25] L. E. Orgel, J. Mol. Bio. **38** (1968) 381.
[26] K. Kruger, *et al.*, Cell **31** (1982) 147.
[27] C. Guerrier-Takada, *et al.*, Cell **35** (1983) 849.
[28] A. C. Forster and R. H. Symons, Cell **49** (1987) 211.
[29] A. Hampel and R. R. Tritz, Biochemistry **28** (1989) 4929.
[30] E. H. Ekland, J. W. Szostak and D. P. Bartel, Science **269** (1995) 364.
[31] S. J. Gould and R. C. Lewontin, Proc. R. Soc. Lond. B **205** (1979) 581.
[32] C. Woese, Proc. Natl. Acad. Sci. USA 95 (1998) 6854.
[33] A. Meyer, Nature **421** (2003) 31.
[34] M. Lynch and L. C. Conery, Science **290** (2000) 1151.
[35] J. Maynard Smith, *Evolution Genetics*, (Oxford University Press, 1998).
[36] S. Otto and P. Yong, Adn. Genetics **46** (2001) 451.
[37] Z. Gu, *et al.*, Nature **421** (2003) 63.